# MATCHING IN EPIDEMIOLOGICAL STUDIES*

**Suharyanto Supardi**
**Division of Public Health**
**Faculty of Medicine Gadjah Mada University**
**Yogyakarta**

## INTRODUCTION

In the assessment of an association between two variables one must make allowance for certain confounding variable that may affect the relationship being studied. Efforts to control confounding variable could be made at the design stage of the study and/or at the analysis stage. There are three relatively simple method available for the control of confounding variable matching, stratification, and covariance or regression control. Matching is to be considered as control of confounding variable at design stage and the rest are at analysis stage. Matching can be done in two ways, individual or pair matching and category or frequency matching. Random sampling method usually to be done in the selection of comparison groups, unexposed subjects in follow-up study and non-diseased subjects in case control study.

Decision on preference of sampling strategy for choosing the comparison subject based on statistical consideration (*i.e.* statistical efficiency) and assumption no relationship between the method of sampling and the size of comparison group.

The objectives of this paper are as follows :

1. To compare between category matching and pair matching based on statistical consideration.
2. To discuss the advantages and disadvantages of category matching,to compare between category matching and pair matching based on statistical consideration.
3. To discuss consideration that should be taken into account when deciding the method of sampling strategy and its statistical analysis.

---

*Based on article in American Journal of Epidemiology, Vol. 116, No. 5, entitled, "In Defense of Matching" by John M. Karon and Lawrence L. Kupper, 1982.

**Berita Kedokteran Masyarakat III:6**
**ISSN 0215-1936. Jil.III No.6 Juni 1987**

## DEFINITION OF TERMS

1. Confounding variable (confounder) is a third variable or extraneous variable that could affect in the (crude) estimate as result of mix-up of the exposure-disease relation-ship understudy with the effects of this variable.

2. Matching is procedure to eliminate the effect of variables that may confound the analysis of study variable by which selection of controls is done in such a way (partial restriction) so that the control group has the same distribution as the cases with respect to certain confounding variables (risk factor).

3. Individual matching is a paired sampling with additional requirement that the individuals selected from the control population that much match the corresponding case with respect to specified criteria or characteristics.

4. Frequency/category/stratified matching is a procedure in which samples are selected from every subgroup that has been subdivided according to the chosen variables, and different proportions - corresponding to the distributions in the case series.

5. Statistical efficiency is a evaluation of statistical procedure, e.g. sampling procedure, that reflected in precision in effect measure estimation and power in hypothesis testing.

### Category Matching vs Pair Matching

Performing a matched pair is justifiable only when the number of matching variable is large enough (or the categorization is refined enough) so that most, of all, of the individual strata will contain only one index subject and one comparison subject. In other words, pair matching will be relevant when the categories are so refined that there is just one index subject and one comparison subject in each stratum.

It should be taken into account since in many studies, especially case-control studies, pair matching has been used to control confounder and reduce subject variability. Pair matching is usually carried out by forming categories even for continuous

variables. This situation would result in loss of statistical efficiency. Therefore, pair matching is inferior to category matching and there is no statistical basis for pair matching on a categorical variable since pair is not unique.

That means that category matching gives a more precise estimate than does pair matching. It can be seen in comparison of the Montel-Haensael $X^2$ statistic for category matched data and the McNemar $X^2$ for pair matched data. The reason of it is that the pair is not unique, hence using pair matching will result in overmatching.

In practise, no attempt is made to check the interaction when one is using pair matching, since only overall effect measure is computed. Therefore, it is possible of obscuring a much different association in the strata.

Category matching must be considered in choosing the comparison subjects instead of pair matching unless in very specified condition.

**Advantages and Disadvantages of Category Matching**

Some advantages of category matching are as follows :

1. Efficient adjustment for the potential confounding effects of wide range of social-economic factors that would difficult to control through matching on variables like sibship or neighbourhood of residence.

   Matching on that variable would reduce the variability over strata with regard to the ratio of the number of comparison subjects to the number of index subjects, especially when dealing with small sample size. (The variability in the matching ratio can lead to inefficient analysis and even to completely uninformative strata).

2. Matching saves money and time.

3. Matching on risk factor that be considered apriory to be highly likely to manifest as strong confounders in the data, and still retain the flexibility to adjust if necessary, for other factors at the analysis stage.

4. Category matching gives a more statistically efficient analysis than choosing the same number of reference by random sampling. Efficiency is expected more pronounced in follow-up studies than in case-control studies. And also the stronger the risk factor being matched on, the larger will be the gain in efficiency over random sampling.

**Some disadvantages of category matching are as follows:**

1. Matching can result on information loss due to discarding of available reference not able to satisfy the matching criteria. Time consuming and labor required process to fing appropriate matches. If the reference loss is more than 50% then random sampling can be more statistically efficient method.

2. Matching prevents evaluation over the underlying population relationship between the matching variable and exposure status or disease status in its respective study design.

3. Category matching in case-control studies can sometimes lead to a loss in efficiency relative to random sampling. It depends on the strength of relationship between matching factor with disease variable and exposure variable in population.

Therefore, matching procedure can be useful in follow-up and case-control studies of prior knowledge about important relationships among the disease, exposure and all potentially confounding variables is available and is used properly. If it doubt, the safest strategy is to match only on strong risk factors expected to be differentially distributed between the exposed and unexposed study subjects.

Another considerations should be taken into account when considering matching on one or more extraneous factor are cost and logistics.

**Comparison Category Matching and Random Sampling**

1. In follow-up study, below is an illustrative example.

Table 1. Follow-up Study to Estimate Risk Expected Cell Counts

a. Unexposed chosen by random sampling from the population

| | $F_0$ | | | $F_1$ | | | TOTAL | | |
|---|---|---|---|---|---|---|---|---|---|
| | E | $\bar{E}$ | TOTAL | E | $\bar{E}$ | TOTAL | E | $\bar{E}$ | TOTAL |
| D | 8 | 6 | 14 | 64 | 8 | 72 | 72 | 14 | 86 |
| $\bar{D}$ | 192 | 594 | 786 | 736 | 392 | 1128 | 928 | 986 | 1914 |
| Total | 200 | 600 | 800 | 800 | 400 | 1200 | 1000 | 1000 | 2000 |

b. Unexposed chosen by Category Matching on f

| | FO | | | F1 | | | TOTAL | | |
|---|---|---|---|---|---|---|---|---|---|
| | E | $\bar{E}$ | TOTAL | E | $\bar{E}$ | TOTAL | E | $\bar{E}$ | TOTAL |
| D | 8 | 2 | 10 | 64 | 16 | 80 | 72 | 18 | 90 |
| $\bar{D}$ | 192 | 198 | 390 | 736 | 784 | 1520 | 928 | 982 | 1910 |
| Total | 200 | 200 | 400 | 800 | 800 | 1600 | 1000 | 1000 | 2000 |

Table 2. Statistics Related to the Estimation of Risk Ratio

| Choice | Risk Ratio Estimates | | | | 95 % Confindens | | |
|---|---|---|---|---|---|---|---|
| Unexpo. | Crude | Fo | F1 | dRR | Intervals for a | $PR^D$ | $x^2$ values$^D$ |
| Random | 5.14 | 4.00 | 4.00 | 4.00 | (2.20 , 7,26) | | 23,83 |
| Sampling | 4.00 | 4.00 | 4.00 | 4.00 | (2.14 , 6.65) | | 33.98 |

1) 45% confidence interval for RR is computed using Taylor series-based weights is of the general form (a $\widehat{RR}$) exp ($\pm 1.96/Wo + W1$).

2) $X^2$ is computed using Mantel-Haenzel $X^2$ statistics. Some interpretations can be drawn from this example.

1) Category matching gives more precise estimate of population RR than random sampling. In general true if sample size chosen in category matching not less than 50% of the sample size chosen by random sampling.

2) Category matching on f has ensured that confounding variable is not present.

2. In Case-control study ; a numerical example.

Table 3. Case-control Study to Estimate Exposure Odds Ratio Expected Cell Counts

a. Non-diseased chosen by random sampling

| | $F_0$ | | | $F_1$ | | | TOTAL | | |
|---|---|---|---|---|---|---|---|---|---|
| | E | $\bar{E}$ | TOTAL | E | $\bar{E}$ | TOTAL | E | $\bar{E}$ | TOTAL |
| D | 16 | 48 | 64 | 128 | 64 | 192 | 144 | 112 | 256 |
| $\bar{D}$ | 10 | 125 | 135 | 39 | 82 | 121 | 49 | 207 | 256 |
| Total | 226 | 173 | 199 | 167 | 146 | 313 | 193 | 319 | 512 |

b. Non-diseased chosen by Category Matching on f

| | $F_0$ | | | $F_1$ | | | TOTAL | | |
|---|---|---|---|---|---|---|---|---|---|
| | E | $\bar{E}$ | TOTAL | E | $\bar{E}$ | TOTAL | E | $\bar{E}$ | TOTAL |
| D | 16 | 48 | 64 | 128 | 64 | 192 | 144 | 122 | 256 |
| $\bar{D}$ | 5 | 59 | 64 | 61 | 131 | 192 | 66 | 190 | 256 |
| Total | 221 | 107 | 128 | 189 | 195 | 384 | 210 | 302 | 512 |

Tabel 4. Statistics Related to the Estimation of Exposure Odds Ratio

| Choice of non-diseased. | Exp.Odds Ratio Estim. Crude$_{OR}$ | $F_1$ | F | $\widehat{aEOR}$ | 95 % Confidence Intervals for aEOR | $X^2$ values |
|---|---|---|---|---|---|---|
| Random Sampling | 5.43 | 4.17 | 4.20 | 1.20 | (2.75 , 6.40) | 46.92 |
| Category Sampling | 3.70 | 3.93 | 4.30 | 4.24 | (2.85 , 6.31) | 53.41 |

Some interpretations can be drawn from this example.

1. Category matching in case-control study has not controlled confounding variables therefore, stratified analysis still be needed at analysis stage to control confounding variable.

   This is so because category matching in confounding variable does not affect the relationship between exposure and confounding variables in the sample, meaning its relationship still reflects the population corresponding relationship.

2. Category matching in case-control study gives more efficient analysis than random sampling does. Therefore, it is necessary to match confounding variables that strongly related to disease variable (risk factor) and are expected to be differentially distributed between the exposed and unexposed groups when the control subjects are chosen by random sampling.
3. Cost consideration must be taken into account when relative efficiency are to be concluded, since category matching can result in a reduction in the number of control subjects.

   Some explanation should be given with regards why category matching on a single dichotomous variables controls confounding variables with respect to the risk ratio in follow-up studies but not with respect to the odds ratio in case-control studies. In follow-up studies category matching ensure there would be no relationship between exposure and the confounder in the sample, OR ef = 1, therefore there is

no effect modification in the population, hence we expect that the crude $\widehat{RR}$ to be approximately equal to common stratum-spesific value. While in case-control studies category matching ensures that there will be no relation ship between disease and confounder in the sample, $\widehat{OR}$ df = 1, but it is not sufficient for controlling confounding. Therefore, stratified analysis would be needed to control confounder.

Table 5. Recommendations for Analysis Based on Validity and Precision by Study Design, Population Association and Method of Sampling

| Study Design | Presence of Confounding and Population Associations | Selection of Comparison Group | |
|---|---|---|---|
| | | Category Matching | Random Sampling |
| Follow-up (using RR) | | | |
| | No: $OR_{df} = 1$, $OR_{df/\bar{E}} = 1$ | Stratified[a] | Stratified[a] |
| | No: $OR_{df/\bar{E}} = 1$ | Unstratified[b] | Unstratified |
| | Yes: $OR_{ef/\bar{D}} \neq 1$, $OR_{df/\bar{E}} = 1$ | Unstratified[b] | Unstratified |
| Case-control (using OR) | | | |
| | No: $OR_{ef/\bar{D}} = 1$ | Unstratified[b] | Unstratified |
| | No: $OR_{ef/\bar{D}} \neq 1$, $OR_{df/\bar{E}} = 1$ | Stratified | Unstratified |
| | Yes: $OR_{ef/\bar{D}} \neq 1$, $OR_{df/\bar{E}} \neq 1$ | Stratified | Stratified |

[a] The unstratified data provide a valid point estimate of the RR. A stratified analysis can be more efficient.

[b] The unstratified and stratified data yield the same estimate forward: OR. However, the Mantel-Haenzel chi-square for the stratified data is slightly smaller than that for the unstratified data.

## Efficiency

The choice of unexposed and non diseased subjects in follow-up or case-control study are selected either by random sampling or by category matching on third factor will be based on

statistical efficiency consideration. It means that which method of sample selection gives the more efficient estimate.

The following are some general recommendation :

1. Every known risk factor that expected to have substantially different distributions in the exposed and unexposed group should be matched.

2. In category matching the chosen comparison subjects could be found at more expensive cost. It if is so therefore, the composition subjects would be smaller than if random sampling is used

3. Matching may reduce the sample size in a study since it is required that comparison group should have the same distribution over strata as the index group.

4. If good estimates are available for the odds ratios describing some of the relations among disease, exposure and confounding factor than the decision in choosing between matching and random sampling can be made.

In follow-up matching the considerations are as follows :

1. Matching should be done when matching factors is confounder in population and has sufficient strong relation with exposure (ORef1).

2. Any loss in efficiency due to matching will be relatively small, if there is no loss of subjects due to matching.

3. A gain in efficiency could be achieved by matching on risk factors and if there is strong relation between exposure and the potential confounder.

4. A meaningful gain in efficiency by using random sampling can be achieved if the unexposed subject has two or three times as large as under category matching.

Hence, it can be concluded that category matching should be considered in follow-up studies, particularly in small and medium size studies.

In case-control studies, the following are some considerations:

1. If there is no relation between exposure variable and matching factor among nondiseased, $(OR)_{ef/\bar{D}} = 1$, choosing the controls by category matching and random sampling, gives equivalent results.

2. Matching on strong risk factor will give gain in efficiency. If there is no risk factor, random sampling will give gain in efficiency as $OR_{ef/\bar{D}}$ increases. Therefore, matching should only be done on risk factor.

3. Random sampling will give meaningful gain in efficiency when number of subjects in control groups chosen by this method have at least twice as many as chosen by matching; whether or not f is confounder.

4. Matching is more efficient than random sampling when matching factor is risk factor, the number of control subjects is not affected by sampling method, even if there is statistical interaction.
   It can be summarized in Table 5.

## CONCLUSION

1. When category matching on a confounder is used to select subjects.

   a. In follow-up study on unstratified analysis gives a valid point estimate of the risk ratio.
   b. In case-control study a stratified analysis is necessary to obtain valid point estimate of the odds ratio.

2. Stratification at analysis stage can be used when the control group is chosen by random sampling given information can be obtained.

3. Adjustment for potential confounder of socio-economic factor, which is very difficult to measure, by stafily on a surrogate character such as sibship or neighbourhood or residence, random sampling should not be used.

4. Decision on whether or not to match based on statistical efficiency which are dependent on matching factor should be risk factor and effect of matching on sample size of

comparison group.

5. a. Matching will give more efficiency in follow-up study if it has at least 75% comparison's sample size as many as many as the study using random sampling and confounder present in the population.

   b. In case-control study the situation is similar when at least 40-50% of control subjects compare to random sampling and strong relationship between exposure variable and confounder.

6. Pair matching is not the method of choice for controlling confounding but rather post stratification should be used in the situation are have considered.
   Its summary can be seen in Table 6.

Table 6. Summary of Conclusions About Efficiency Based on Comparison in Table 5 (Cost Considerations Ignored)

| Study Type | No Population Confounding | Population -Confounding |
|---|---|---|
| Follow-up | $OR_{ef}=1$ or $OR_{df/\bar{E}} = 1$; no expected gain or loss from matching | $OR_{ef}\neq 1$ and $OR_{df/\bar{E}} \neq 1$ ; expected gain from matching |
| Case-control | $OR_{ef/\bar{D}}=1$ ; no expected gain or loss from matching<br>$OR_{ef/\bar{D}}\neq 1$ and $OR_{df/\bar{E}} = 1$ ; expected loss from matching when EOR and exposure rates are larges | $OR_{ef}\neq 1$ and $OR_{df/\bar{E}} \neq 1$ ; expected gain from matching<br>$OR_{ef/\bar{D}}\neq 1$ and $OR_{df/\bar{E}} \neq 1$; expected gain from matching when EOR and exposure ratesaresmall to moderate;expected loss from matching when EOR and exposure rates are large. |

REFERENCES

Currie JB. 1971 Matched and Unmatched : A Comparison of Two Designs, with Epidemiologic Data. , *American Journal of Epidemiology* ; 93 : 315-316.

Fisher L. and Patil K., 1974 Matching and Unrelatedness, *American Journal of Epidemiology* ; 100 : 347-349.

Fleiss J.L., 1981 *Statistical Methods for Rates and Proportions,* second editions, John Wiley and Sons, Inc. New York.

Kleinbaum, Kupper, and Morgenstern, 1982 *Epidemiologic Research Principles and Quantitative Methods*, Lifetime Learning Publications, pp. 377-401.

MacMahon and Pugh TF. , 1970 *Epidemiology Principles and Methods,* Little, Brown and Coy., Boston, pp. 235-236.

Miettinen OS., 1970 Matching and Design Efficiency in Retrospective Studies, *American Journal of Epidemiology;* 91:111-118.

---

**Berbagai Pandangan untuk SBS-2000**

*Dalam menghadapi konsep **SEHAT BAGI SEMUA DI TAHUN 2000** orang-orang yang optimis berharap pada tahun 2000 nanti semua orang akan menikmati hidup sehat. Kalangan pesimis berolok sehat untuk 2000 orang pada tahun 2000, dan yang lain lagi berseloroh semua orang baru akan menikmati hidup sehat 2000 tahun lagi.*

---